ISSS617: Python for Data Science

PROJECT PROPOSAL

Group G1-6

Akash Khowala

Amit Singla

Neel Chomal

Pankaj Bhootra

Wu Yinxiao

Topic: Trump Twitter Analysis

We understand what plagiarism is and have ensured we did not plagiarise for this assignment. This assignment is in partial fulfilment of the requirements for the module ISSS617 Python for Data Science.

1. Introduction

This project focuses on one of the most controversial and influential figures on this planet at the moment: Donald Trump. The current POTUS uses social media to express opinions on world moving issues in respect of politics, military, markets, trade, economy, etc. We believe that Trump's electronic musings have a positive correlation with impacting global issues and as part of this project, we will be looking at the effect of his tweets on financial markets. Research has shown that social media plays a leading role in forecasting events, using thoughts of people by collecting, storing and analysing data for the purpose of harvesting information. Social media analytics is focused on developing frameworks to collect, monitor, analyse, summarize, and visualize social media data to extract useful patterns.

We will analyse all the tweets posted by Trump through various Twitter handles (@POTUS, @realDonaldTrump, and relevant allies) to understand the topics he posts about, the sentiments expressed in the tweet (positive, negative, neutral), run a timeline analysis by comparing the sentiments during a period with the trends in various market indexes (such as treasuries market, stock market, etc.) during the same period and build a predictive model for the same. We have a large corpus of Donald Trump tweets going back several years, along with data on historical market trends, which will give us enough insight to capture his impact and draw relevant conclusions.

2. Scope of Problem and Stakeholders

This project will only cover Trump's potential impact on financial markets and as such, we will run topic modelling algorithms to filter out irrelevant tweets. A potential application of our project could be around building a predictive market index to capture the market impact of Trump's tweets in real-time, for which the stakeholders would include banks, brokerages and investors, who are interested in predicting market trends accurately.

3. Previous Work

2019, Bloomberg Report: https://www.bloomberg.com/news/articles/2019-09-09/jpmorgan-creates-volfefe-index-to-track-trump-tweet-impact

JP Morgan and Citigroup, two of the world's largest banks, have successfully measured the market impact of POTUS' tweets.

JP Morgan's analysts have created the infamous "Volfefe Index" to track this impact on treasury yields. For every five minutes after Trump tweeted something with direct Fed references, the Volfefe Index showed "a rolling one-month probability that each missive is market moving." Specifically, they found that short term securities were more impacted than 10-year ones.

Citi has similarly shown that Trump's tweets are generally followed by volatile behaviour across currency markets globally. They have forecasted that markets are likely to show more sensitivity to Trump's continued tweeting with USD and Fed references.

4. Datasets Used

No.	Description	Dataset Link for Reference
1	Twitter dataset provided by Professor	https://drive.google.com/drive/folders/1r5HNcKU-gQzjPFvTMk8wtSTWXzGmPvW7
2	Kaggle Tweets dataset	https://www.kaggle.com/austinreese/trump-tweets
3	Market Trends, eg. US Treasury dataset	https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield

We will be looking at multiple dataset sources to consolidate Trump's tweets for the past several years. We will be using the Twitter dataset provided as part of this course and filter it for tweets posted by "@realDonaldTrump", "@POTUS" and other allied accounts.

Additionally, we will use the Kaggle dataset above, which has tweets posted only by "@realDonaldTrump" between May 2009 and January 2020.

The above datasets have the following relevant information available:

1) User ID – Unique identifier for the tweet posted by a user in Twitter database
2) Content – The tweet text for analysis
3) Timestamp – The date tweet was posted by the user
4) Hashtags – Hashtags (if any) used in the tweet
5) Geolocation of tweet (if specified) – Location the tweet was posted from

In addition to procuring Trump's historical tweets, we are also interested in gathering datasets corresponding to trends in various markets, such as treasury yields, stock exchanges, currency markets, etc. For instance, we have found a dataset for US treasury posted by the US Department of Treasury (link on previous page). This dataset is in XML format and provides daily treasury yield curve rates (historical and projected). We also aim to procure other relevant datasets to get enough information on market trends for past several years.

5. Analytical Methodologies to be Explored

- **Topic Analysis** – Topic analysis is a Natural Language Processing (NLP) technique that allows us to automatically extract meaning from texts by identifying recurrent themes or topics. Finding out key words, for example – jobs, trade, military, etc. using clustering to study the impact of Trump's tweets on the respective industry/sector.
- **Sentiment Analysis** – Contextual mining of text which identifies and extracts subjective information in source material and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. We will be segregating the tweets on 'positive', 'negative' and 'neutral'. This will help us generate data for our prediction analysis.

- **Exploratory Data Analysis** (EDA) – A critical step for performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. The objective here is to find out the variables that have an impact on the market trends (e.g. treasury yields) for the analysis we will be doing later in the project.
- **Hypothesis Testing** – A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is testable based on observing a process that is modelled via a set of random variables. For example, in case of treasury yields:
Recognizing correlation between movement in treasury interest rates over time as compared to the sentiment of the Trump's tweets.

  Yes – Trump's hypothesis have an impact on the treasury yield curve rates.

  No – Trump's tweets do not have an impact on the treasury yield curve rates.
- **Predictive Analysis** – Predictive analytics encompasses a variety of statistical techniques such as predictive modelling, and machine learning, that analyse current and historical facts to make predictions about future or otherwise unknown events. We aim to build a predictive model to incorporate all the above analysis and predict market trend given a Trump tweet.

6. Expected Results

Our expectation is to find a positive correlation between Trump's tweets and trends in specific markets. We use previous research by analysts at JP Morgan and Citigroup as our basis for this expectation as they have successfully shown Trump's impact in specific markets. Our objective will be to establish certain definitive findings around the ramifications of Trump's market takes which could potentially be of use in designing systemic strategies.

7. Project Timeline

A Gantt chart is provided below to illustrate project milestones and individuals accountable for each outcome. This progress is tracked across 8 weeks starting from February 9th. The final completion is marked for April 19th (1 week before final report deadline).

Task/Week	1	2	3	4	5	6	7	8	Akash	Amit	Neel	Pankaj	Yin
Ideation, feasibility study													
Proposal work													
Dataset preparation													
EDA and PCA													
Predictive modelling, insights and recommendations													
Final Report and Presentation													

8. Conclusion

The development of effective and efficient analytics techniques for social media analysis is of utmost importance. To conduct social media analytics, data mining, text analysis and related advanced techniques such as sentiment analysis and semantic analysis are frequently adopted. There is a growing interest in the power of social media analytics in creating new value, supporting decision making and enhancing competitive advantage. A number of studies from various research communities have been devoted to unleash the value, impact and implications of social media analytics.

With Donald Trump's tweets, we have a splendid example of social analytics use case that can potential provide accurate estimation for market trends. We are excited about this problem and look forward to discovering what patterns emerge from our analysis of Trump's tweets.